Estimating the Rate of Overdiagnosis With Prostate Cancer Screening: Evidence From the Finnish Component of The European Randomized Study of Screening for Prostate Cancer

SD Walter (walter@mcmaster.ca)
McMaster University
Jiarui Hu
McMaster University
Kirsi Talala
Cancer Research UK
Teuvo Tammela
Tampere University
Kimmo Taari
University of Helsinki
Anssi Auvinen
Tampereen Yliopisto

Research Article

Keywords: Prostate cancer, screening, PSA, randomised trial, over-diagnosis, mortality.

Posted Date: March 31st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-341289/v1

Version of Record: A version of this preprint was published at Cancer Causes & Control on July 27th, 2021. See the published version at https://doi.org/10.1007/s10552-021-01480-8.

Abstract

Purpose: Screening for prostate cancer may have limited impact on decreasing prostate cancer-related mortality. A major disadvantage is overdiagnosis, whereby lesions are identified that would not have become evident during the man's lifetime if screening had not taken place. The present study aims to estimate the rate of overdiagnosis using Finnish data from the European randomized trial of prostate screening.

Methods: We used data from 80,149 men randomized to a screening or a control group, distinguishing four birth cohorts. We used the "catch-up method" to identify when the difference in the cumulative incidence of prostate cancer between the screening and control groups had stabilised, implying that the screening has no further effect. We define the overdiagnosis rate to be the relative excess cumulative incidence in the screened group at that point. As an independent method, we also examined the diagnosis rates of T1c tumours as an indicator of early tumors detected by PSA.

Results: The estimates of overdiagnosis rates from the catch-up method using the full period of available follow-up ranged between cohorts from 2.3% to 15.4%, and the T1c analysis gave very similar results.

Conclusions: Some overdiagnosis has occurred, but there is uncertainty about its extent. A long follow-up is required to demonstrate the full impact of screening. We evaluated the overdiagnosis rates at a population level, associated with being offered screening, taking account of contamination (screening among the controls). The overall evaluation of screening should incorporate mortality benefit, cost-effectiveness and quality of life.

Introduction

Prostate cancer is the second commonest cancer in males worldwide, but different regions have varying incidence and mortality. The risk of prostate cancer is higher in black men but is low in Asian men [1]. In the US, the most commonly diagnosed cancer in men is in the prostate. The American Cancer Society [2] estimated that during 2018, about 164,690 new cases of prostate cancer would be diagnosed in the US.. In Canada, it has been estimated that about 21,300 men would be diagnosed with prostate cancer annually.

In 1986, PSA testing was approved by the Food and Drug Administration (FDA) to monitor the progression of prostate cancer. In 1994, the FDA approved the use of PSA in screening for prostate cancer in asymptomatic men. As a result, the incidence rates for prostate cancer increased substantially in the 1980s and 1990s, primarily because of widespread adoption of the PSA test. However, a recent analysis showed that the incidence of distant-stage prostate cancer increased among men ages 50 to 69 between 2004 and 2012 [4]. Moreover, an ACS guideline updated in 2001 indicated there was still uncertainty about the overall value of periodic testing in terms of reducing the risk of death from prostate cancer. A randomized trial conducted in the US found no mortality benefit [5], whereas a contemporaneous trial conducted in eight countries in Europe showed a 21% reduction in prostate cancer mortality [6]. At that

point, the ACS recommended that PSA testing was not recommended for asymptomatic men who had less than a 10-year life expectancy, and physicians were required to provide detailed information to their patients about the risks and potential harms of early detection. Also, a large cluster-randomised trial in the UK showed no mortality benefit at 10 years, but it was based on a single screening round with low compliance (36%) [3].

The European Association of Urology recommends the provision of PSA testing to informed men with elevated risk of prostate cancer, with follow-up intervals for men depending on their initial PSA levels. In 2017, the US Preventive Services Task Force recommended that men aged 55-69 should be informed about the benefits and harms of PSA-testing, in order to decrease the number of men with aggressive disease being missed. All these findings imply that systematic population-based PSA-testing is not strongly recommended [7].

While the benefits of PSA testing remain controversial, there has also been concern about the adverse effects of PSA testing, particularly with respect to the question of overdiagnosis. Specifically, PSA testing may detect some cancers which would not have been identified during a man's lifetime had screening not taken place [8]; the diagnosis of such lesions through screening clearly provide no mortality benefit. Such overdiagnosis could result from the presence of slow growing or indolent tumors, which can exist asymptomatically for many years. In these cases, screening potentially leads to harmful effects, such as erectile dysfunction, urinary incontinence and others. However, at the time of screening, it is impossible to recognize which particular cases of cancer have been over-diagnosed and should have been left untreated. Even for aggressive cancers, it is possible that men will die before the cancer has time to progress; in such cases, this would also amount to over-diagnosis. Accordingly, in order to evaluate the overall benefits or harms of prostate screening, we need to quantify the extent of overdiagnosis in a screening program.

Two main approaches have been suggested to estimate the overdiagnosis rate: modeling of disease transition rates, and the "catch-up" or excess-incidence method [9]. The first approach models the hypothetical or counterfactual patterns in prostate cancer that would arise with or without screening, then comparing their results to estimate the rate of overdiagnosis. Examples of this method include: MISCAN [8, 10] which is a microsimulation model that simulates individual life history as a Markov process of states and transitions to calculate the over-detection rate by deriving the lead time; the UMich (University of Michigan) method [11], in which a statistical model captures the features of registered prostate cancer cases before and after PSA screening was used, then predicts lead time and subsequently the overdiagnosis rate; and the FHCRC (Fred Hutchinson Cancer Research Center) method [12, 13]) in which a microsimulation model links an individual's PSA levels with the progression of his prostate cancer.

In all of these simulation methods, investigators have to find a balance between complexity and transparency in choosing an appropriate model. The complexity dimension can range from simple (involving only a few features of the disease) to complex (referring to many disease features and adopting many transitional probabilities). If a complex model is used, it may be difficult to evaluate the

risk of bias in the results, due to a lack of transparency. On the other hand, a simple model may not capture all the important features of the disease process, and its interface with screening.

The second approach, the so-called “catch-up” method, uses observed excess incidence rates, and the cumulative difference in disease incidence between the screening and the control groups. In a review of alternative approaches to assessing overdiagnosis, the catch-up method has been described as the preferred approach [14], and that it is particularly applicable to situations where randomized trial data are available, such as in the Finnish data employed in the present analysis. Taking advantage of the data from a randomized control group allows one to more reliably estimate the expected disease pattern in the counterfactual scenario where screening has not been used.

In this study, the catch-up method was used in the Finnish component of the European Randomized study of Screening for Prostate Cancer (ERSPC) [15]. In this trial, men were individually randomized to be offered PSA testing (the screened group), or to a control group where screening was not offered. By virtue of the randomization, we can assume that the two groups have the same underlying risk of prostate cancer. We estimated the extent of cancer overdiagnosis by examining the pattern over time in the cumulative difference in the incidence rate of prostate cancer diagnosis between the screening and the control groups, during the follow-up period after the end of the screening intervention. We used regression methods to assess the point during the follow-up period when the difference in the cumulative incidence for all prostate cancer diagnoses had stabilised, indicating that the impact of the screening intervention had worn off. We also verified the results using a separate analysis of stage T1c tumours (defined as early, clinically inapparent, non-palpable cancers). The estimates of overdiagnosis rates reflect the comparison between the intervention and control groups as a whole, so in other words they evaluate the effect of being offered to participate in a screening program or not. As such, any PSA testing that occurs in either group outside the trial itself is taken into account.

Methods

Data was abstracted from the Finland section of the ERSPC, which is a multi-center, randomized screening trial between an intervention arm offered PSA screening and a control arm without an intervention. In Finland, one of eight participating countries, 80458 men aged 55-67 years were randomized to a screening or a control arm, distinguishing four birth cohorts: 1941-44,1937-40, 1933-36 and 1929-32. Men in the three youngest cohorts in the screening group were offered up to three rounds of prostate screening at four-year intervals, in 1996-99, 2000-2003, and 2004-2007; the final round excluded men aged >71 years; men in the oldest cohort were offered only two rounds of testing, starting in the same year. A PSA level 4.0 ng/ml was used as the indication for biopsy. For men with PSA between 3.0 ng/ml and 3.99 ng/ml a digital rectal examination was initially offered as a supplementary (reflex) test in 1996-1998, and since 1999, free/total PSA ratio was used (with a cut-off of 0.16). In this paper, data obtained during follow-up of trial participants was used for 18.6 years after randomisation.

Figure 1 shows a schematic representation of the expected patterns of cumulative incidence of prostate cancer in the screened and control arms of a randomized trial, in either the absence or presence or overdiagnosis. Before screening begins, both arms accumulate cases at the same expected rate. However, during periods of screening, cases are found in the screening arm earlier than would otherwise have occurred; the degree to which the date of diagnosis is advanced is known as the lead time. The earlier distribution of diagnosis dates in the screened group manifests as a difference in cumulative incidence in that group, relative to controls. The difference may be further enhanced during later rounds of screening. When the screening program ends, cases are then diagnosed more frequently in the controls than in the screened arm, because the pool of cases in the screened arm has been somewhat exhausted, and the controls experience their diagnoses later than in the study arm. In the absence of over-diagnosis, as in Figure 1a, one expects all the control counterparts of the screened cases (with early diagnosis dates) to eventually be diagnosed at a later time. After some time, the screening effect will have dissipated, and the cumulative incidence in the controls will “catch up” with that in the screening arm.

In contrast, if there has been over-diagnosis of some cases in the screened arm, their expected control counterparts are never diagnosed, and consequently the cumulative incidence in the control arm always lags behind that of the screening arm. Conceptually, at some point, the cumulative difference in incidence between the screening and control arms will stabilize, and at that stage the cumulative difference will represent the number of over-diagnosed cases in the screened arm (Figure 1b). We define the estimated overdiagnosis rate as the cumulative difference in incidence at this “stability point”, divided by the cumulative incidence in the screened group.

The challenge is to determine when (or if) catch-up has occurred. We modelled the differences in the year-specific incidence rates with spline regressions. Using year-specific incidence, rather than the cumulative incidence difference, has the advantage that the incidence data points are mutually independent. We attempted to determine the stability point by identifying when the slope of the year-specific rate differences was at or close to zero. Our initial impression was that some of the trends in the Finnish data were not clear-cut, and that it might therefore be empirically difficult to define when stability had occurred. Accordingly, we also evaluated the performance of the spline regression method with simulated data.

In the simulations, the year-specific incidence rates of each cohort were assumed to follow a Poisson distribution, which could be approximated by a Normal distribution. We assumed that the ideal pattern of incidence rate differences for the spline regressions would demonstrate patterns approximately as shown in Figure 2, with the time axis starting at the end of the screening program. In the model of Figure 2, there are up to three linear segments (or splines), with two join points (or ‘breakpoints’). The sharp initial decrease occurs because of the early depletion of the pool of cases in the screened group. Then, for some period of time, the screening arm accrues cases at a lower rate than the controls. Finally, as the screening effect wears off, the control incidence rate converges to and eventually equals the rate in the screened group, and catch-up is then declared to have occurred. The rate difference at the catch-up point is zero, and hence we would conclude that there had been no over-diagnosis. However, if the rate difference stabilizes at a non-zero value, that value will provide the estimated extent of over-diagnosis.

The idealized model in Figure 2a can be fitted if there are a sufficient number of data points for each spline segment, and if the follow-up period after the end of screening is long enough to actually observe stability in the rate difference, once the effect of screening has dissipated. If the data were insufficient to fit this model, a compromise two-segment model was adopted, as in Figure 2b, in which there is not enough data to distinguish the second and third segments of the model in Fig 2a. If the follow-up appears to have ended ends before stability of the rate difference can be identified, then a simpler model with only two splines and one join point was adopted, eliminating the final segment in the model of Figure 2b, after the stability point.

In the simulations, we repeatedly fit the various spline regressions, to evaluate the performance of that method. We sampled the distributions of the year-specific incidence rates in the screening and control arms, based on the numbers of detected prostate cancer cases and the numbers of men at risk in each study year. The variance for each distribution was taken to be the same as the empirical mean rate, assuming Poisson distributions for the numbers of cases.

Because the rates in the screening and control groups are statistically independent, the variances of the rate differences can be taken as the sum of the two group-specific variances. Then, by appealing to the Central Limit Theorem, the year-specific rate differences were assumed to approximately follow a Normal distribution with this combined variance. For each simulated sampled of data points, we attempted to fit the spline regression, and thus to estimate the catch-up point. Each simulation scenario was initially repeated 100 times, but if the number of converged regression fits was less than 50, we increased the number of simulation runs to 200, to acquire sufficient converged solutions with a specified number of spline segments. The final estimate of each parameter was taken as the sample mean calculated from the simulated set of fitted spline regressions if the distribution of the parameter was symmetric, but otherwise the median was used.

Because the alternative spline models with different numbers of component segments are not hierarchical, the Akaike Information Criterion (AIC) was used to select the number of breakpoints required. The AIC provides a way to consider the trade-off between the goodness-of-fit of each model to the data and its complexity. The model chosen by this criterion then gives estimates of the times of each breakpoint, and the slopes of each spline segment (to be denoted as slope1, slope 2, and slope 3, as appropriate).

In the three-segment models (as in Figure 2a), the first break point was conceptualized to be when the year-specific rate differences had reached their lowest point, and the second breakpoint is when the rate difference has become stable, and it was taken as the catch-up point. Initial values of the breakpoints, which are required for the iterative fitting of the spline regressions, were based on visual impressions of the plotted data.

A prerequisite for data in which there is a well-defined “catch-up” point is that there are enough years of follow-up, which ideally needs to be at least as long as the longest lead time that screening can provide [16]. For prostate cancer, the mean lead-time has been estimated as between about 5 and 8 years in

various analyses and populations [8, 17, 18]. Thus, the available follow-up time of over 18 years since randomisation likely exceeds the lead time for most cases. However, to the extent that catch-up has still not fully occurred, there will be some tendency to overestimate the overdiagnosis rate.

We defined overdiagnosis to be the detection of cancers by screening that would not have become clinically evident in the absence of screening. In situations where the catch-up point could be identified, we estimated the overdiagnosis rate as (-)/ , where and are the cumulative incidence rates in the screened and control groups, respectively, at the catch-up point. A 95% confidence interval for the rate of overdiagnosis was calculated as where and are the standard errors of the corresponding cumulative rate differences.

In addition to examining the cumulative incidence of all prostate cancer diagnoses, we also carried out a separate analysis of T1c tumours, which are typically asymptomatic. The empirical values of the difference in the cumulative incidence of these tumours were compared to the catch-up estimates of overdiagnosis at the latest points during the follow-up. The T1c analysis will reflect PSA testing both within the trial and outside it, as by definition a T1c cancer is a clinically inapparent tumour that is not palpable in digital rectal examination or visible in imaging (but not an incidental finding in transurethral resection of the prostate as T1a and T1b); it is frequently detected because of an elevated PSA as it is too small to cause symptoms.

It is important to recognize that our estimates of overdiagnosis rates reflect comparisons between the intervention and control groups as a whole; so in other words they evaluate the effect of being offered to participate in a screening program or not. As such, any PSA testing that occurs in either group outside the trial protocol itself is taken into account, including ‘contamination’ testing (screening or symptom-driven) of men in the control group.

Results

Prostate cancer Incidence

Data used in this study was taken from the Finland data in the ERSPC, conducted in men born from 1929 to 1944. A total of 80,458 men were randomized to screening or control groups. Table 1 shows the sample sizes and distribution of follow-up times available for the 1929-32, 1933-36, 1937-1940, and 1941-1944 cohorts; all men are followed indefinitely, until death, or individuals were censored once a prostate diagnosis had occurred. Figure 3 shows their cumulative incidence, the year-specific incidence rate, and their differences. Immediately evident is the fact that the cumulative incidence is progressively higher for the earlier birth cohorts, as would be expected [19]; accordingly, all our analyses were done separately for each cohort. The cumulative incidence plots do appear to support our initial conceptualization for their expected behavior, as displayed in Figure 1.

The data for the 1929-32 cohort (Figure 3a) appears to approach a zero cumulative difference between the screening and control groups, while the other cohorts retain non-zero differences. The two peaks in

year-specific incidence correspond to the two screening rounds in the study protocol (during years 1 and 5 of follow-up) for this cohort. After the end of screening in follow-up year 5, the screened group incidence fell below the controls because of the lead-time effect, and then the groups gradually converged at a catch-up point of about 16 to almost 19 years of follow-up since randomisation.

In the three later birth cohorts, there are three years of excess incidence in the screening group corresponding to their screening protocol, followed by a deficit after year 9. The deficit continues for several years, then the screening group incidence gradually returns to that of the controls (Figures 3b, 3c, and 3d).

Figure 3 also shows the cumulative excess incidence rates. In each cohort, the cumulative excess achieves its maximum value at the time of the last screening round. None of the cohorts clearly attain a zero cumulative incidence difference by the end of follow-up, suggesting that some over-diagnosis may be present in each case, but that the effect of screening may persist beyond the last year of follow-up.

Model selection

Table 2 shows the AIC statistic for the various spline regression models in each cohort; smaller values suggest the preferred model, among the cases where convergence of the model fitting was successful. On this basis, the appropriate numbers of breakpoints were defined as 1 for the 1929-32 and 1933-36 cohorts, and 2 for the 1937-40 and 1941-44 cohorts.

1929-32 cohort:

Based on the AIC statistic, the preferred model for this cohort has one break point, at the point where the rate difference has its lowest value.

Among the 100 simulation runs, 98 converged for the spline model with 1 joint point; summaries of the model parameters are shown in Table 3. The point when the year-specific rate difference reached its minimum was at 2.29 years. The estimated slope of the second segment was small, but zero was not contained within its whiskers *(max(Q1-1.5*(Q3-Q1), min), min(Q3+1.5(Q3-Q1), max))* [20], (this range is approximately $\mu \pm 2.67\sigma$ under a normal distribution assumption) which suggested that it was significantly greater than 0.

Figure 4a shows the fitted two-segment model to the observed data. It has a minimum around the second year of follow-up, which is close to the mean value in the simulated samples, and has a subsequent to rise to approximately 0. We conclude that either ‘catch-up’ may have occurred, but there is insufficient data to define a later breakpoint after which the year-specific rate differences would have completely stabilized at zero.

1933-36 cohort:

We adopted the two-segment spline model. All the simulation runs converged, and their estimated parameters are again summarized in Table 3. The minimum rate difference was approximately at 2.4 years after the last screen.

Figure 4b shows two-segment model fitting to the observed data with a minimum incidence difference estimated at about 2.5 years of follow-up, but with a slow upward trend after that. The last few years of follow-up show variable incidence rate differences, both above and below zero, so again it is not completely clear if the catch-up point has been reached.

1937-40 cohort

We used a three-segment spline model with two break points. In order to acquire a larger sample of converged simulations, the number of replications was increased from 100 to 200; the results are summarized in Table 3, for the 80 simulations (40%) which converged. Non-convergence often occurred because there was only one data point in some time segments, or because two breakpoints were close to each other.

The distributions of the estimated slopes showed positive skewness for breakpoint 1, and negative skewness for breakpoint 2, so we adopted the mean values as the preferred summary, because qualitatively these values were close to their corresponding median. The minimum difference in year-specific incidence rates was reached just over two years after screening ends, then there is a slowly increasing trend until about 8 years. The mean slope of the third segment was positive over the short period of remaining follow-up data available.

The three-segment model fitted to the observed data is shown in Figure 4c, indicating a rapid drop in the cumulative incidence rate difference for the first two years, and then a period of about 7 years with an approximately stable deficit in negative values; an increase is seen in the last year of available data, suggesting that a stable catch-up point may not yet have occurred.

1941-44 cohort

The pattern of year-specific rate differences for this cohort was similar to that of the 1937-40 cohort. In this case, about 70 (30%) of the three-segment model simulations converged, with non-convergence again occurring when there was only one data point in one or more segments or two closely-spaced breakpoints. Mean values were used to estimate breakpoints and slopes, because they were close to their corresponding medians in all cases.

The three-segment spline regression models fitted to the cohort data are displayed in Figure 4d. The small difference between slope2 and slope3 illustrates the difficulty of identifying the time of the second break point, and this also explains why the standard deviation of joint point 2 is much larger than for joint point 1. Once again, we could not definitively identify if catch-up had occurred.

Estimated over-diagnosis rates

Table 4 shows estimates of the absolute and relative overdiagnosis rate, based on the cumulative incidence difference between the screened and control groups, for various periods of time since the last screen. The absolute cumulative incidence rate difference (i.e. the cumulative excess risk of prostate cancer) for men born in 1929-32 was 0.004 (95% confidence interval: -0.011,0.019) at 14 years since the end of screening. Compared to the cumulative incidence in the screening group, the relative overdiagnosis rate was therefore (0.004 / 0.176) * 100% = 2.3%. This indicates that 2.3% of men who started screening at age 67-70, and who were identified by the screen to have prostate cancer, were over-diagnosed.

For men who started screening at age 63-66, 59-62 and 56-58, the cumulative incidence differences after 10 years of follow-up after the last screen were 0.026, 0.015, and 0.010. The corresponding relative rates of over-diagnosis were 15.4%, 11.4% and 10.2%, respectively. This suggests proportionally greater absolute differences in incidence among older men, and with correspondingly higher rates of over-diagnosis, in these three cohorts, who each had three screens offered. However, the oldest cohort (born 1929-32), which was offered only two screens, does not reflect this trend.

A difficulty in interpreting these estimates of overdiagnosis is that PSA testing has occurred in the control group of the ERSPC, and also in the intervention group outside the regimen of the trial itself. Furthermore, a PSA test is used in the diagnostic process for almost all cases of prostate cancer. Finally, it is not possible to say, from the available data, whether some of these tests were true screens in asymptomatic men, and which tests might have been administered in response to symptoms, i.e. for clinical indications. As noted elsewhere, testing within the control group would probably tend to cause under-estimation of over-diagnosis. Despite this, it is not possible to devise a correction for this effect, because of the uncertainties surrounding the motivation for particular tests. In response to this concern, we carried out an additional analysis of diagnosis rates for prostate cancer T1c tumors, which are defined as clinically inapparent tumours that are not palpable nor detected in surgery for benign prostatic hyperplasia (transurethral resection of the prostate). This means that most early tumors detected by PSA testing would be classified as T1c.

We constructed life tables for T1c diagnoses in both arms of the trial, again with censoring when a prostate cancer diagnosis or death had occurred. From the cumulative incidence rates, we calculated relative overdiagnosis rates in the last year of follow-up data. The relative overdiagnosis rates based on T1c diagnoses for the 1929-32, 1933-36, 1937-40 and 1941-44 birth cohorts were 2.3%, 16.3%, 14.6% and 12.7% respectively, agreeing very closely with the estimates from the catch-up method using all prostate diagnoses, which were 2.3%, 15.4%, 11.4% and 10.3%. This supports the notion that the catch-up estimates are valid, in the sense of allowing for all tests in all men in the trial, and with the objective of estimating over-diagnosis in the trial groups as a whole.

Discussion

Based on our results, we found that the available years of follow-up (over 18 years since randomisation, and 10 years after the end of their last screening round) in the three youngest cohorts in the trial were not

quite enough for us to definitively confirm whether the incidence difference between the screening and control groups had stabilized or not; the oldest 1929-33 cohort, with 14 years of follow-up after the end of their last screening round, shows somewhat more convincing evidence that catch-up of the control group had occurred. Elsewhere, it has been estimated that 10-14 years of follow-up may be required [21]. It is possible that the cumulative incidence for prostate cancer will continue to reduce with further observation. If so, the best available estimate of the overdiagnosis rate would be calculated from the data in the last year of follow-up, but this would be an overestimate if the screen effect is still wearing off, even at that late stage.

Table 5 summarizes the estimates of overdiagnosis obtained in other studies; these range widely, from 2.9% to 88.1%. Such substantial variation might be partly explained by the fact that in deriving these estimates, there are many possible choices for the denominator [9]. Studies by Etzioni et al. [22], Telesca et al. [17], and ourselves report overdiagnosis as a percentage of screening-detected cases. Others presented the number over-diagnosed as a proportion of the total number of cases detected, or the total number invited to screening. The variation may also be attributed to different methodologies being employed. In most modeling studies, investigators used disease incidence rates in a screening group to estimate the distribution of the lead time, or to infer natural history of the disease, and subsequently estimate the corresponding frequency of overdiagnosis.

In contrast, an approach based on observing the excess incidence in a screened group leads more directly to an estimate of the overdiagnosis rate. However, there are two challenges: first, how to estimate the incidence in unscreened persons; and second, a requirement to have sufficient follow-up years where the protocol of the trial is respected. Concerning the former, Zappa et al. [27] estimated the incidence without screening based on the pre-screening trend, while Schröder et al. [28] used data from a randomized clinical trial. Concerning the latter challenge, having a long period of follow-up may make it difficult to avoid men in the control arm from being screened during the study years. Therefore, when possible, the screening contamination rate of the control group should also be considered. Also, in the Finland data, approximately 10% of men in the screening group had had a PSA test before their first screen in the trial [31], (although, being pre-randomisation, these tests will have been equally distributed between the study arms). More recently, it has been estimated that 50% of the control men in Finland have had a screening test at least once during the first eight years of follow-up [32]. Such a high contamination rate in the control arm will tend to reduce the excess incidence between the two groups, and thus lead to an underestimate of the overdiagnosis rate, if even the follow-up is long enough for the incidence difference to become stable. Nevertheless, our analysis has had the advantage of estimating overdiagnosis compared to a randomized control group that was not offered screening as part of the trial. Other approaches, such as comparing outcomes in screened individuals with the pre-screening trend, do not have the obvious benefits of randomization. In addition to differences in their analytic methods, further reasons for the variation between the results of studies summarized in Table 5 include differences in the screening protocols and techniques.

The fact that some PSA testing also took place in the controls is an important factor in the interpretation of our results. Because of the way the testing data was accessed for the community-based control men, we do not know if any given PSA test in that group was carried out as a true screen (asymptomatically), as opposed to symptom-driven testing. Furthermore, PSA testing is involved in the process of making almost all prostate cancer diagnoses, including in the intervention group, and again we cannot tell which particular tests should "count" as screens in either group. There will surely have been some 'contamination' of the control group by true screening tests, and although PSA testing among the controls was quite frequent [33], we cannot say how often this occurred as true screens. The same is true of the intervention group. Because of these uncertainties, it is not possible to 'correct' or adjust for non-screening PSA tests carried out on the trial participants. Such an adjustment, if it were possible, could lead to estimates of the prostate cancer diagnosis rates among individual men actually screened *versus* a counterfactual scenario where screening did not take place, in other words as an efficacy comparison, but not one whose validity is protected by the randomisation.

However, it is important to recognize that our estimates of overdiagnosis rates reflect comparisons between the intervention and control groups as a whole, i.e. in an effectiveness context. In other words, they evaluate the randomized comparison of groups being offered to participate in a screening program or not. As such, any PSA testing that occurs within either group but outside the trial protocol itself is taken into account. Overdiagnosis can actually occur in individual men within either group, but our randomized comparisons reveal the difference between the overdiagnosis rates for the entire groups in an unbiased way.

We have therefore carried out the new analysis of T1c tumours, which are defined as early, clinically inapparent and not detectable by digital rectal examination or transrectal ultrasound, which leaves PSA as the likely indication for a prostate biopsy. The results were very consistent with the main catch-up analysis of this paper. It should be noted that all our estimates of over-diagnosis rates apply, in the first place, to the population studied in the Finnish component of the ERSPC. The importance of this problem elsewhere will potentially vary according to factors such as the *ad hoc* PSA testing behaviour by asymptomatic men, the distribution of risk factors for prostate cancer, and patterns of other morbidity.

In conclusion, we have examined the feasibility of using regression modelling to find the "catch-up" point when the effect of screening has worn off, and the cumulative incidence difference between screened and control men has become stable. Based on the Finland data, we concluded that the cumulative incidence difference at the last available year of follow-up may have led to some over-estimation of the overdiagnosis rate. Our best estimates of the relative overdiagnosis rate were 2.3%, 15.4%, 11.4%, and 10.3% for the 1929-32, 1933-36, 1937-40, and 1941-44 cohorts, respectively. Theory suggests that the overdiagnosis rate might increase with age, because of the combined effects of a higher detection rate and higher rates of other causes of death in older men [34]. However, the lower over-diagnosis rate for the oldest men in our results could be explained by the fact that there were only two screening rounds for the 1929-32 cohort. Also, recall that we estimate over-diagnosis rates based on the entire experience of each study arm, including PSA tests in either arm that may or may not be associated with the screening trial

itself. Finally, note that we have estimated rates of relative overdiagnosis; further work on this topic might consider absolute rates of overdiagnosis, and contrast them against the NND (the number needed to detect), i.e. evaluate the number of over-detected cases versus one averted death.

Declarations

***Funding**:* This paper was partially funded by the Natural Sciences and Engineering Research Council, Canada; the Academy of Finland (Grant No 260931); the Cancer Foundation Finland; and Competitive State Research Funding administered by Pirkanmaa University Hospital special responsibility area (grants 9E089, 9H099, 9F100, 9R002).

Conflicts of interest/Competing interests

All authors declare they have no conflict of interest, with the exception of Dr Auvinen who has received a lecture honorarium from Novartis.

Ethics approval (include appropriate approvals or waivers); An ethics committee review was conducted by Tampere University Hospital (tracking no 95077; R10167). All men who participated in screening gave written consent. For the control arm, registry-based follow-up without contact with study participants does not require consent, according to the Finnish regulations.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and material: Finnish privacy regulations and General Data Protection Regulations of the EU do not allow sharing pseudonymised data with sensitive content such as health information.

Code availability: Not applicable.

Authors' contributions All authors contributed to the study conception and design. Material preparation and data collection were performed by Hu, Talala, Tammela, Taari, and Auvinen, and analyses were performed by Hu and Walter. The first draft of the manuscript was written by Walter and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. Culp MB, Soerjomataram I, Efstathiou JA, Bray F, Jemal A (2020). Recent global patterns in prostate cancer incidence and mortality rates. Eur Urol 77: 38-52.

2. American Cancer Society. (2018). Cancer Facts & Figures 2017. Atlanta: American Cancer Society.

3. Martin RM, Donovan JL, Turner EL, et al. (2018). Effect of a low-intensity PSA-based screening intervention on prostate cancer mortality: the CAP randomized clinical trial. *JAMA* 319: 883-895.

4. Hoffman RM, Meisner ALW, Arap W, Barry M, Shah SK, Zeliadt SB, Wiggins CL (2016). Trends in United States prostate cancer incidence rates by age and stage, 1995–2012. Cancer Epidem Biomar 25: 259-263.

5. Andriole GL, et al., for the PLCO Project Team (2009). Mortality results from a randomized prostate cancer screening trial. New Engl J Med360:1310-1319.

6. Hugosson J, Roobol MJ, Mansson M, Tammela TLJ, Zappa M, Nelen V, Kwiatkowski M, Auvinen A (2019). A 16-yr follow-up of the European randomized study of screening for prostate cancer. Eur Urol 76: 43-51.

7. Mottet N, Bellmunt J, Briers E, et al. (2018). EAU-ESTRO-ESUR-SIOG Guidelines on prostate cancer. Eur Urol 73(5): e134-e135.

8. Draisma G, Etzioni R, Tsodikov A, Mariotto A, Wever E, Gulati R, et al. (2009). Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. J National Cancer Institute101: 374-383.

9. Etzioni R, Gulati R, Mallinger L, & Mandelblatt J (2013). Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. Ann Intern Med 158: 831.

10. Draisma G, De Koning H (2003). MISCAN: Estimating lead-time and over-detection by simulation. BJUrol Int92(s2):106-111

11. Tsodikov A, Szabo A, Wegelin J (2006). A population model of prostate cancer incidence. Stat Med.25: 2846-2866.

12. Etzioni R, Tsodikov A, Mariotto A, Szabo A, Falcon S, Wegelin J, Feuer E (2008). Quantifying the role of PSA screening in the US prostate cancer mortality decline. Cancer Causes and Control 19: 175–181.

13. Gulati R, Inoue LY, Gore JL, Katcher J, Etzioni R (2014). Individualized estimates of overdiagnosis in screen-detected prostate cancer. J National Cancer Institute 106(2):djt367.

14. Biesheuvel C, Barratt A, Howard K, Houssami N, Irwig L (2007). Effects of study methods and biases on estimates of invasive breast cancer over-detection with mammography screening: a systematic review. Lancet Oncol 8:1129-1138.

15. Schröder, F H, Hugosson J, Roobol MJ, Tammela TLJ Zappa M, Nelen V, ERSPC Investigators. (2014). Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. Lancet 384: 2027-35.

16. Gulati R, Feuer EJ, Etzioni R (2016). Conditions for valid empirical estimates of cancer overdiagnosis in randomized trials and population studies. Am J Epidemiol 184(2):140–147.

17. Telesca D, Etzioni R, Gulati R (2008). Estimating lead time and overdiagnosis associated with PSA screening from prostate cancer incidence trends Biometrics 64, 10–19.

18. Finne P, Fallah M, Hakama M, Ciatto S, Hugosson J, Koning HD, Auvinen A. (2010). Lead-time in the European randomised study of screening for prostate cancer. Eur J Cancer 46(17):3102-3108.

19. Bell K, Del Mar C, Wright G, Dickinson J, Glasziou P (2015). Prevalence of incidental prostate cancer: A systematic review of autopsy studies. Int J Cancer 137(7):1749-1757.

20. Krzywinski M, Altman N. (2014). Visualizing samples with box plots. Nature Methods11:119-120.

21. Auvinen A, Määttänen L, Stenman U, Tammela T, Rannikko S, Aro J, Juusela H, Hakama M (2002). Lead-time in prostate cancer screening (Finland). Cancer Cause Control13:279-285.

22. Etzioni R, Penson DF, Legler JM, Di Tommaso D, Boer R, Gann PH, Feuer EJ (2002). Overdiagnosis due to prostate-specific antigen screening: lessons from US prostate cancer incidence trends. J Natl Cancer Inst94(13):981-990.

23. Wu GHM, Auvinen A, Maattanen L, Tammela TLJ, Stenman UH, Hakama M, et al. (2012). Number of screens for over-detection as an indicator of absolute risk of overdiagnosis in prostate cancer screening. Int J Cancer 131:1367-75.

24. Pathirana T, Hayen A, Doust J, et al. (2019). Lifetime risk of prostate cancer overdiagnosis in Australia: quantifying the risk of overdiagnosis associated with prostate cancer screening in using a novel lifetime risk approach. BMJ Open9:e022457.

25. Gulati R, Morgan TM, A'mar T, Psutka SP, Tosoian JJ, Etzioni R (2020). Overdiagnosis and lives saved by reflex testing men with intermediate prostate-specific antigen levels, *J National Cancer Institute 112: 384 – 390.*

26. Gulati R, Psutka SP, Etzioni R. (2019). Personalized risks of overdiagnosis for screen-detected prostate cancer incorporating patient comorbidities: Estimation and communication. J Urology.202(5):936-943.

27. Zappa M, Ciatto S, Bonardi R, & Mazzotta A (1998). Overdiagnosis of prostate carcinoma by screening: An estimate based on the results of the Florence Screening Pilot Study. Ann Oncol9(12):1297-1300

28. Schröder F, Hugosson J, Roobol M, Tammela T, Ciatto,S, Nelen V et al. (2009). Screening and prostate-cancer mortality in a randomized European study. New Engl J Med 360(13):1320-1328

29. Ciatto S, Gervasi G, Bonardi R, Frullini P, Zendron P, Lombardi C, Zappa M (2005). Determining overdiagnosis by screening with DRE/TRUS or PSA (Florence pilot studies, 1991–1994). Eur J Cancer41(3):411-415.

30. Fenton JJ, Weyrich MS, Durbin S, Liu Y, Bang H, Melnikow (2018). Prostate-specific antigen–based screening for prostate cancer: evidence report and systematic review for the US preventive services task force. JAMA 319(18):1914-1931.

31. Ciatto S, Zappa M, Villers A, Paez A, Otto SJ, Auvinen A (2004). Contamination by opportunistic screening in the European randomized study of prostate cancer screening. British J Urology 92(s2):97-100.

32. Nevalainen J, Stenman U, Tammela TL, Roobol M, Carlsson S, Talala K, Auvinen A (2017). What explains the differences between centers in the European screening trial? A simulation study. Cancer Epidemiol 46:14-19.

33. Kilpelainen TP, Pogodin-Hannolainen D, Kemppainen K, Talala K, Raitanen J, Taari K, Kujala P, Tammela TLJ, Auvinen A (2017). Estimate of opportunistic Prostate Specific Antigen testing in the Finnish Randomized Study of Screening for Prostate Cancer. J Urolology 198: 50 – 57.

34. Pashayan N, Duffy S, Pharoah P, Greenberg D, Donovan J, Martin R, et al. (2009). Mean sojourn time, overdiagnosis and reduction in advanced stage prostate cancer due to screening with PSA: implications of sojourn time on screening. Brit J Cancer100(7):1198-1204.

Tables

Table 1: Sample sizes and distribution of available follow-up times available, by birth cohort.

		Follow-up time (years)		
Birth Cohort	Sample size*	25% quartile	Median	75% quartile
1929-32	15500	7.00	14.08	16.58
1933-36	17339	9.20	15.58	16.70
1937-40	21101	11.97	15.66	17.58
1941-44	26426	15.12	16.58	17.58
Total	80366	10.78	15.58	17.58

Table 2 AIC values for spline regression models with two or three segments.				
AIC	**1929-32**	**1933-36**	**1937-40**	**1941-44**
Two-Segment Model	-135.2	-84.6	-100.4	**-109.2**
Three-Segment Model	-132.7	-81.1	-105.7	**-115.7**
AIC : Akaike information criterion				

Table 3: Estimated parameters from spline regression models.

1929-32 cohort; 2-segment model

Parameter	Joint Point(t1)	Slope(b1)	Slope(b2)
Mean	2.29	-0.018	0.0006
SD	0.14	0.00302	0.00019
NC	2%		

1933-36 cohort; 2-segment model

Parameter	Joint Point(t1)	Slope(b1)	Slope(b2)
Mean	2.38	-0.0164	0.00105
SD	0.229	0.00315	0.0003
NC	0%		

1937-40 cohort; 3-segment model.

Parameter	Joint Point(t1)	Joint Point(t2)	Slope(b1)	Slope(b2)	Slope(b3)
Mean	2.14	7.78	-0.016	0.000103	0.00241
SD	0.119	1.413	0.0025	0.000549	0.00164
NC			60%		

1941-44 cohort; 3-segment model.

Parameter	Joint Point(t1)	Joint Point(t2)	Slope(b1)	Slope(b2)	Slope(b3)
Mean	2.3	7.38	-0.01	-0.0001	0.00141
SD	0.217	1.59	0.002	0.000529	0.00157

NC	30%

SD : Standard deviation; NC non-convergence

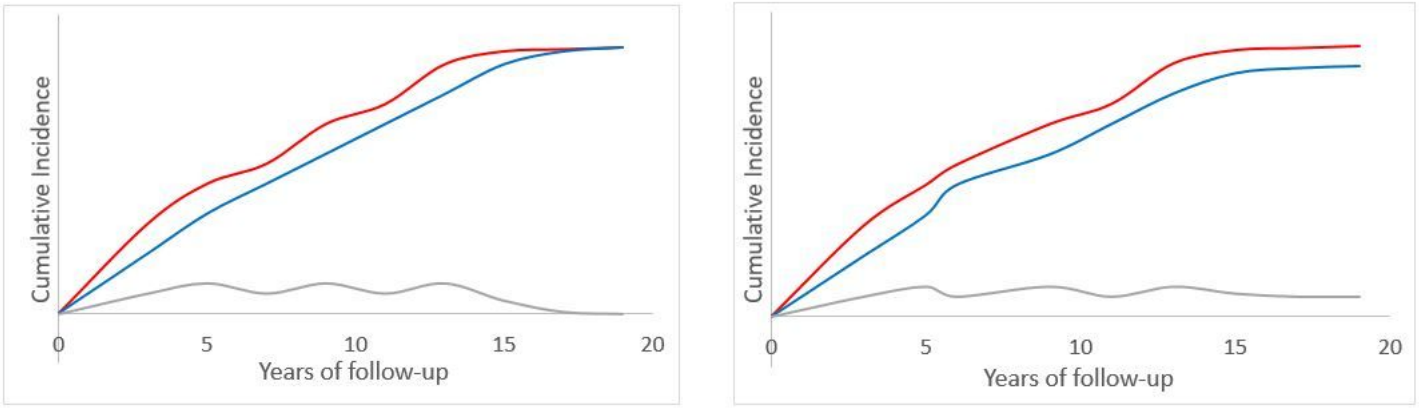
Table 4: Cumulative Incidence excess and estimate of overdiagnosis by birth cohort.

Cohort	Years since end of last screening round	Cumulative Incidence excess of prostate cancer % (Screening versus Control) (95%CI)	Estimated relative overdiagnosis rate (%)
1929-32	12	0.5 (-1.0,2.0)	3.0
	13	0.2 (-1.2,1.6)	1.2
	14	0.4 (-1.1,1.9)	2.3
1933-36	8	2.1 (0.9, 3.3)	13.4
	9	2.6 (1.3, 3.9)	15.7
	10	2.6 1.3, 3.9)	15.4
1937-40	8	1.6 (0.7, 2.5)	13.2
	9	1.3 (0.3 ,2.3)	10.2
	10	1.5 (0.4,2.6)	11.4
1941-44	8	1.2 (0.5,1.9)	14.0
	9	1.0 0.2,1.7)	11.1
	10	1.0 (0.2, 1.8)	10.3

Table 5: Summary of modeling and excess-incidence studies estimating the overdiagnosis rate from PSA testing.

Approach	Author	Study Years	Data	Estimated overdiagnosis rate
Modelling Study	Draisma [10]	2003	ERSPC Rotterdam	48%
	Etzioni [22]	1988-1998	U.S. SEER9	29% in whites, 44% in blacks
	Gulati [13]	1975-2005	U.S. SEER9	2.9-88.1%
	Telesca [17]	1975–2000	U.S. SEER9	22.7% in whites, 34.4% in blacks
	Wu [23]	1996-2005	ERSPC Finland	3.4%
	Pathirana [24]	1982-2012	Australian Cancer Database	41%
	Gulati [25]		10 US clinics	8.8%-60.6%.
	Gulati [26]			4 – 78%
Excess Incidence	Zappa [27]	1992-1995	Italy	51% 25% for 2% annual incremental incidence
	Schröder [28]	1991-2006	ERSPC	48 cases among 1410 screened men
	Ciatto [29]	1991-1994	Italy	66%
	Fenton [30]		ERSPC PLCO CAP trial	33.2% 16.4% 40.7%

Figures

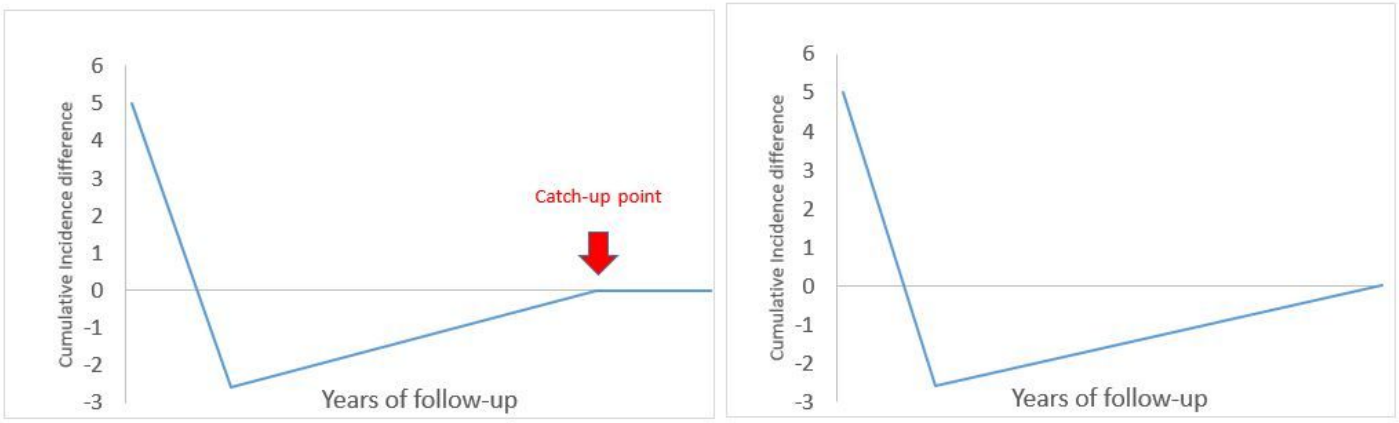


a) Overdiagnosis absent

b) Overdiagnosis present

Figure 1

Schematic expected prostate cancer cumulative incidence patterns in men offered screening (red lines) controls (blue lines), and the difference (grey lines) in (a) the absence, and (b) the presence of overdiagnosis

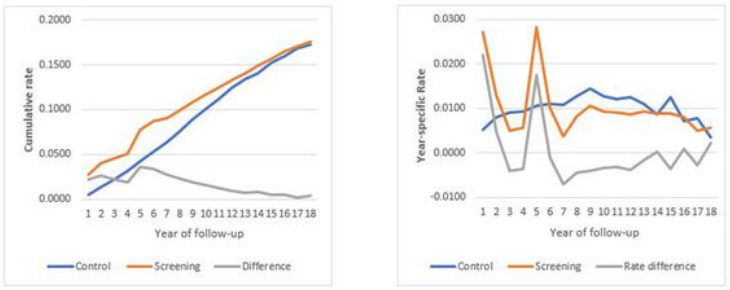


a) Three segment model

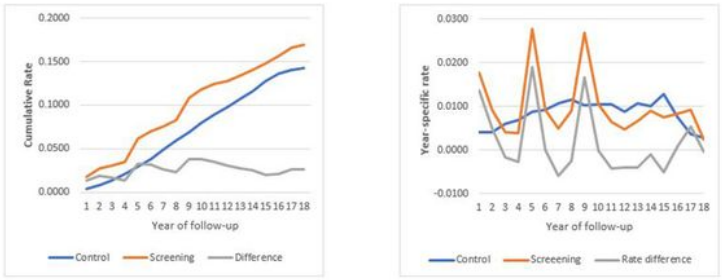
b) Two segment model

Figure 2

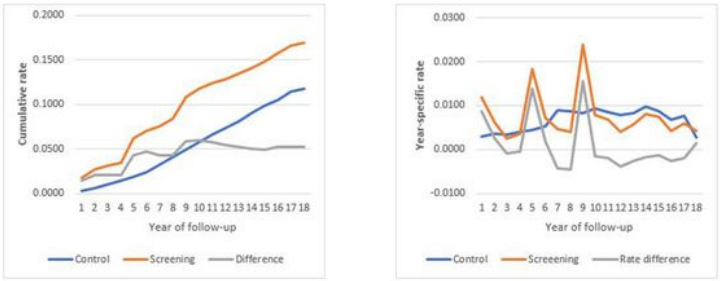
Schematic spline regression models with a) three or b) two segments for the year-specific incidence rate differences.



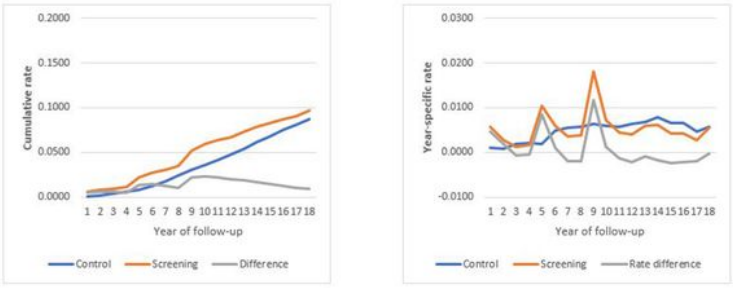
a) 1929-32 cohort



b) 1933-36 cohort



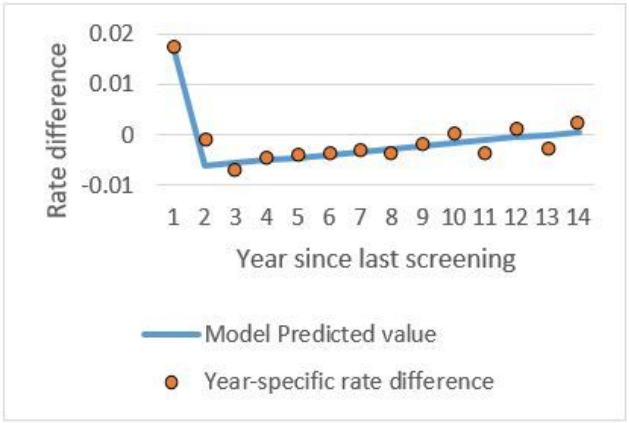
c) 1937-40 cohort



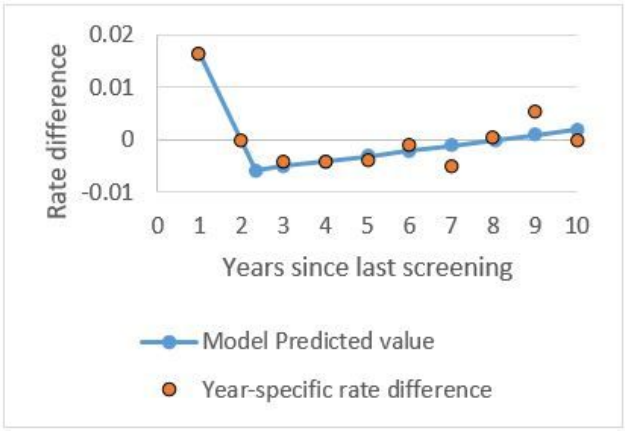
d) 1940-44 cohort

Figure 3

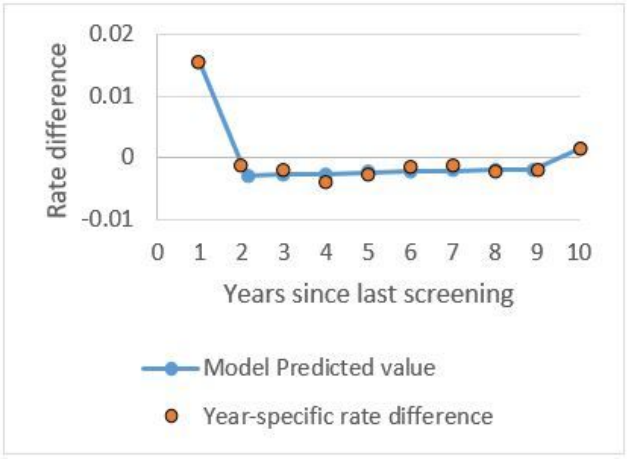
Prostate cancer cumulative and year-specific incidence rates in screening and control arms



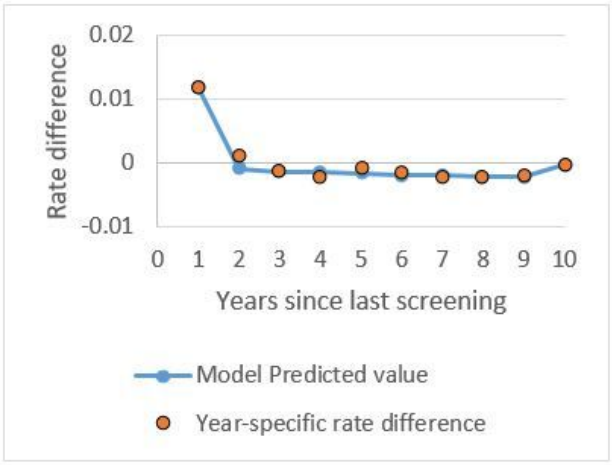
a) 1929-32 cohort



b) 1933-36 cohort



c) 1937-40 cohort



d) 1941-44 cohort

Figure 4

Spline regression model fits to year-specific prostate cancer rate differences.